基于SVM的食双星光变曲线自动分类算法*

袁慧宇[1],赵娟[2],戴海峰[2],杨远贵[2*]
(1.淮北师范大学信息学院 安徽 淮北，235000
2.淮北师范大学 安徽 淮北，235000)

摘 要：提出一种基于机器学习的食双星光变曲线自动分类算法。算法首先对数据进行预处理，将食双星光变曲线数据归一化，并通过滤波/插值降低噪声；随后使用快速傅里叶变换提取频率信号作为特征向量；利用特征向量训练支持向量机获得自动分类模型。使用Python实现算法并抓取CALEB和GCVS数据验证，分析特征向量、支持向量机核函数与惩罚系数对分类正确率的影响，优化后所得分类模型正确率达到92.8%（训练集）和89.0%（测试集），最后使用所得分类模型对第3方数据进行分类，正确率为88.8%，结果证明提出的分类算法有效性。

关键词：光变曲线自动分类；支持向量机；食双星；
中图分类号：TP274 文献标识码：A 文章编号：1672-7673(2019)

在信息与计算技术等新兴科技的驱动下，天文研究领域已从传统的单目标观测和手工处理数据转向多目标观测和自动数据处理[1]，大量巡天项目开展为天文学研究提供了海量数据，如ROTSE，ASAS，SuperWASP，MACHO，OGLE，SDSS，LAMOST和Kepler等，由计算机自动完成目标交叉证认[2]、观测、实时数据处理和分析[2]等，获得光谱、测光、周期以及类型等数据。随着数据量的进一步增大，单服务器已难以实时完成数据处理，分布式计算被应用到数据处理中提高处理效率[4]。面对获得的海量天文数据，必须借助支持向量机、神经网络、遗传算法等人工智能算法对数据进行分析和处理，挖掘有用的信息[5]，如基于随机森林方法对SDSS和XMM数据的天体进行分类[6]；基于机器学习方法寻找射电脉冲信号[7]；基于密近双星的Roche 势对双星进行分类等[8]，所有这些标志着天文学研究已迈入了大数据时代。

通过观测获得的食双星光变曲线，可以快速确定其类型，搜寻出具有特殊演化意义的双星系统，为研究一些特殊天体和现象提供了重要的研究窗口。这对丰富和发展双星的研究内容，通过食双星认识星团和星系的形成和演化具有重要的意义。文[9]使用多项式拟合光变曲线，根据拟合后的曲线的主极小和次极小的宽度和深度给出光变曲线类型；文[10-11]使用傅里叶变换提取光变曲线数据的频率特征，根据所得频率特征进行分类，但在算法实现上使用了软件计算的完美光变曲线数据进行参数设置，使用特征量较少，未考虑仪器测试误差、天气原因等引起的数据波动影响，因此仅能完成对光变曲线进行初步分类，不能识别载有特殊天文现象的光变曲线。

本文提出了一种基于支持向量机的食双星光变曲线自动分类算法，以快速傅里叶变换所得的频率信号为特征量，对支持向量机模型进行训练获得能自动分类的模型。

1. 自动分类算法

---

[1] *基金项目：国家自然科学基金（11473009）；安徽高校优秀青年人才支持计划项目（gxyq2018161）；安徽省高校自然科学项目(KJ2017B017)资助.
收稿日期：2018-07-26；修订日期：2018-08-28
作者简介：袁慧宇，男，硕士.研究方向：计算机机器学习.Email:aquayhy@qq.com
通讯作者：杨远贵，男，教授.研究方向：变星的观测和研究. Email:yygcn@163.com

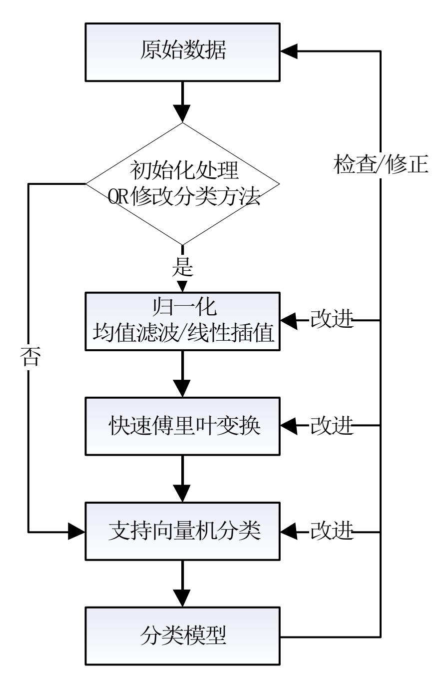


图 1 光变曲线分类方法

Fig.1 Automatic classification algorithm scheme for light curve

食双星光变曲线可分为 EA，EB 和 EW 三种，针对分类需求提出分类方法如图 1，第 1 步对原始数据进行预处理，归一化原始数据并减小其中噪声；第 2 步通过快速傅里叶变换提取频率信号作为特征数据；第 3 步使用基于支持向量机算法训练分类模型并测试；最后对流程优化，获得最优的分类模型。

1.1 数据预处理

ASAS 用理论光变曲线进行分析，无噪声影响[12]，而本文使用 CALEB[2](Catalog and AtLas of Eclipsing Binaries)实测数据（包括相位和较差星等）。由于天气因素以及仪器误差等影响，实测数据不可避免地带有噪声影响。为了降低噪声的影响，首先进行预处理。

（1）归一化，相位数值在[0,1]之间，不需要处理。较差星等可通过式归一化到[0,1]之间。

$$m' = \frac{m - m_{\min}}{m_{\max} - m_{\min}} \quad ,$$

其中，m'为归一化后的较差星等；m 为原始较差星等；m_{max} 和 m_{min} 分别为较差星等最大值和最小值。

（2）使用均值滤波/线性插值算法减少噪声。设 m'' 为预处理后的较差星等最终值。将相位均匀分为 n 段，若第 k 段相位 $\left[\frac{k-1}{n},\frac{k}{n}\right]$ 范围内较差星等的数量 b=1，则采用该数据值作为 m''；若 b＞1，则采用均值滤波算法获得新的 m''，如式；若 b=0，则采用线性插值获得新的 m''，如式。最终获得间隔相等的归一化数据 $\{m''_1, m''_2, \ldots, m''_n\}$。

$$m'' = \frac{1}{b}\sum_{i=1}^{b} m_i \quad ,$$

$$m''_k = \frac{m''_{k-1} + m''_{k+1}}{2} \quad .$$

1.2 光变曲线特征提取

[2] http://caleb.eastern.edu

原始光变曲线为时间序列数据，需将其特征提取出来用于分析。常用特征包括主极小和次极小差值、主极小波谷半高全宽等。本文采用光变曲线的频率特性作为特征值。实际实现时可用快速离散傅里叶变换将相位/较差星等变为频域信号，将频域信号与对应光变曲线类型组成特征数据集$\{f_0,f_1 \cdots f_d,T\}$，其中f为频率分量，T为光变曲线类型。

1.3 支持向量机分类算法

支持向量机是建立在统计学习理论的 VC 维理论和结构风险最小原理基础上的一种有监督的机器学习算法，基本思想是将特征值映射到高维向量空间，获得可将不同类数据分割的超平面，该算法常作为自动分类的机器学习算法。在实际使用中，通常将原始数据分为训练集和测试集。使用训练集训练支持向量机模型，获得映射函数和分割平面（即分类模型），使用测试集验证所得模型。

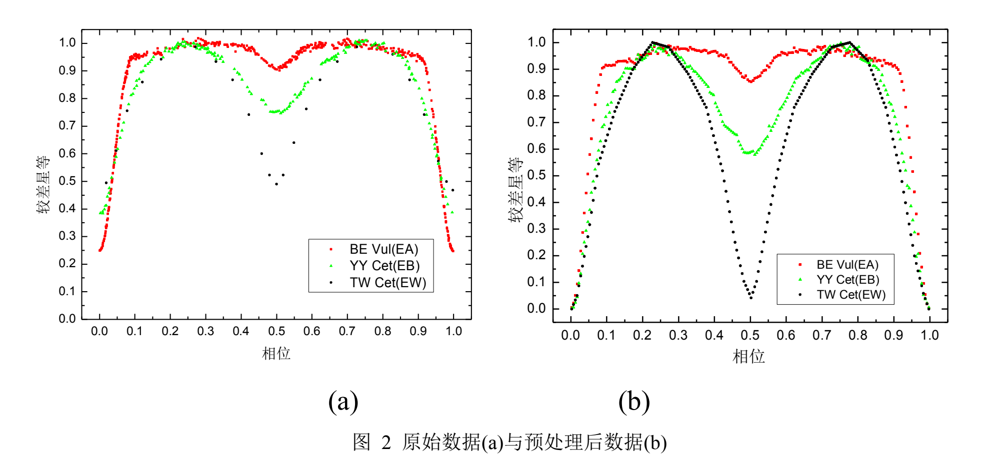
2. 实验与结果分析

算法实现采用 Python 编程，Python 是一种面向对象的解释型计算机编程语言，由于其易用性、简洁和可扩展性，成为最受欢迎的程序设计语言之一。Python 拥有大量的科学计算扩展库在本文算法实现中使用。

2.1 分类实验实现

首先进行原始数据下载和收集， 本文使用 urllib3 和 BeautifulSoup 库自动分析 CALEB 网页数据并存储网站提供的 300 个变星的坐标、星名、类型及 747 条光变曲线，但网站未给出光变曲线类型。随后通过变星坐标与 GCVS[3]（General Catalogue of Variable Stars new version）数据交叉对比获得光变曲线类型。

随后实现光变曲线数据预处理。这里以 BE Vul（EA），YY Cet（EB），TW Cet（EW）三个变星的 V 波段数据为例。原始数据如图 2(a)。由图可知，由于观测设备等限制，观测数据质量较差。表现为数据点个数不一致、浮动较大、数据较离散等。将相位均分为间隔 0.005 的新相位点，应用归一化/均值滤波/线性插值后所得数据如图 2(b)。由图可知预处理保留了原始数据变化趋势，相对原始数据更加平滑。



(a) (b)

图 2 原始数据(a)与预处理后数据(b)

Fig.2 Original data (a) and pre processed data (b)

第三步使用 numpy 和 scipy 库对预处理后数据进行快傅里叶变换完成频域变化。以上文所述 3 颗星数据为例，所得频率值如图 3。其中横坐标代表信号谐波频率。

[3] http://www.sai.msu.su/gcvs/gcvs/intr.htm

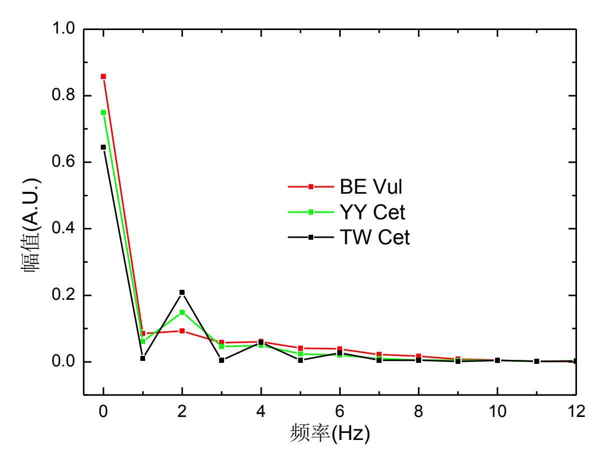
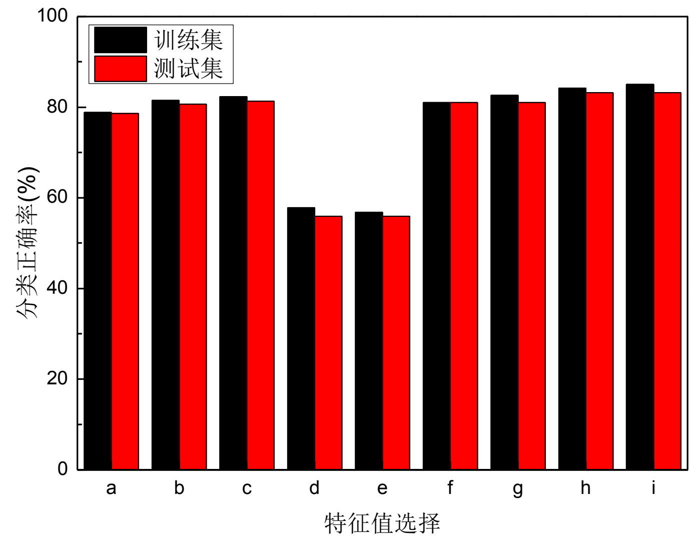


图 3 快速傅里叶变换的结果

Fig.3 Result of FFT

第四步进行支持向量机模型训练。使用上述方法把 747 条光变曲线处理后获得数据集 $\{f_0, f_1 \cdots f_d, T\}$。首先测试频率分量选择对模型训练的影响。用 $[f_i, f_j]$ 表示从 f_i 到 f_j 的连续频率分量集合，用 $\{f_x, f_y\}$ 表示 f_x， f_y 独立的频率分量集合。支持向量机模型选用线性核函数，训练集为 373 条数据，测试集为 374 条数据，惩罚因子设为 1.0。其中核函数是将输入空间映射到高维空间的函数算法。惩罚因子是对错误分类的容忍度，降低容忍度能获得更好的训练结果，但也可能产生过拟合。最终得结果如图 4。由图 4 可知，选取偶次谐波作为特征值时分类正确率较高（图中数据 a、b 和 c），即使仅用 f_0 也可获得 78.6%的分类正确率（图中数据 a）。选择奇次谐波分量作为特征值，正确率最高仅为 57.8%（图中数据 d、e），说明奇次谐波分量不适用于作为特征值。比较图中结果 f 到 i，正确率随着选取的频率数量的增多而上升，说明选择更多频率分量有助于优化分类结果。训练集和测试集正确率相差小于 2%，证明训练结果有效，且未达到过度训练。综合以上结果，偶次谐波分量适用于作为特征值。



a$\{f_0\}$,b$\{f_0, f_2\}$,c$\{f_0, f_2, f_4\}$,d$\{f_1, f_3\}$,e$\{f_1, f_3, f_5\}$,f$[f_0, f_2]$,g$[f_0, f_4]$,h$[f_0, f_6]$,i$[f_0, f_8]$

图 4 分类正确率与特征值关系

Fig.4 Relationship between classification accuracy and characteristic value

2.2 支持向量机优化

接下来优化支持向量机参数设置以获得更好结果，支持向量机参数主要包括核函数选择和惩罚因子设置。选择不同核函数和惩罚因子，使用数据集 $\{f_0, f_2, f_4, f_6, f_8\}$ 作为特征值，最终所得结果如图 5。由图 5 可知 4 种核函数按优劣顺序依次为 linear，rbf，sigmoid 和 poly。

提高惩罚因子初期能显著提升 linear，rbf 和 sigmoid 分类正确率，但在某一阈值后正确率达到稳定。惩罚因子对 poly 无影响。当选用 linear 核函数，惩罚因子设置为 2.0 时，获得的最优分类模型，分类正确率分别为 89.8%（训练集）和 84.8%（测试集），已训练好的模型可以保存，用于其他新的光变曲线数据分类与识别。

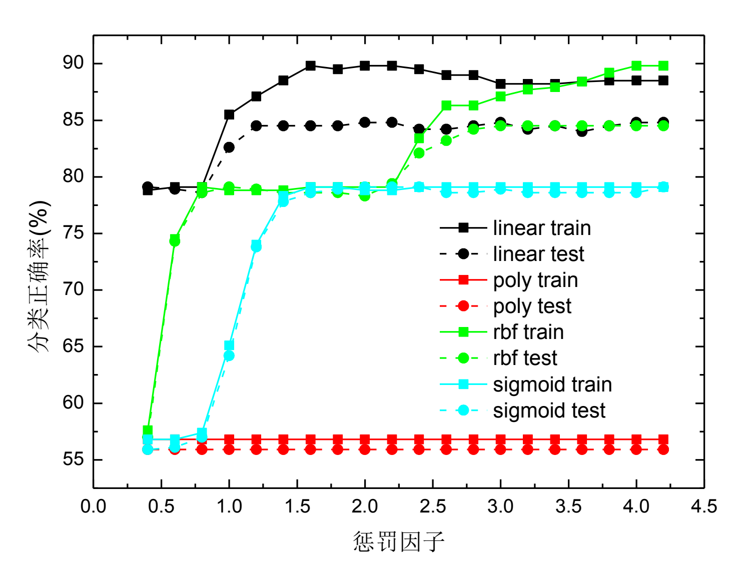
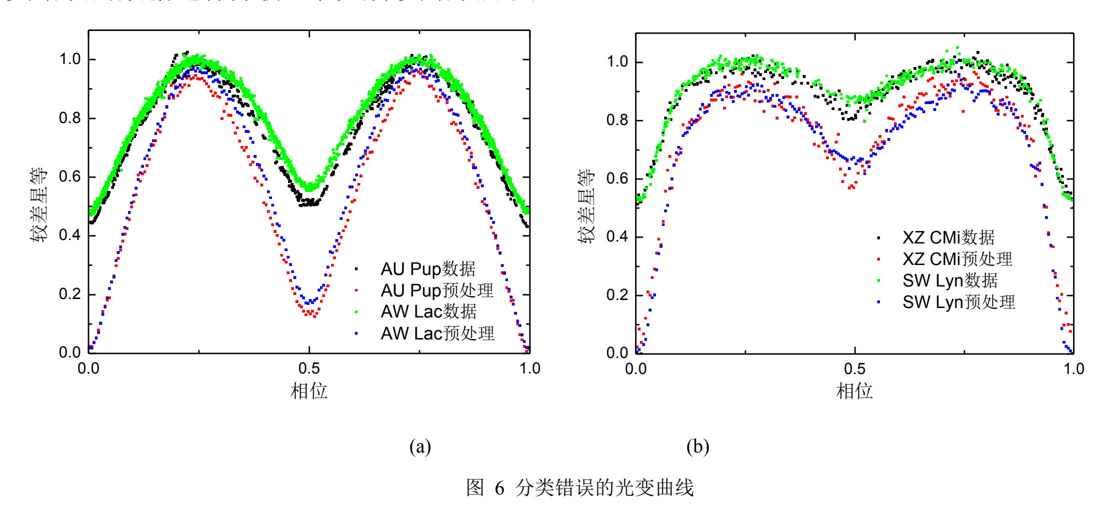


图 5 支持向量机参数与分类正确率关系

Fig.5 Relationship between SVM parameters and classification accuracy

2.3 实验结果分析与数据修正

从结果上看训练所得模型正确率高，能满足分类需求。但仍有分类错误数据，下面对分类错误的数据进行分析，找出分类错误原因。



(a) (b)

图 6 分类错误的光变曲线

Fig.6 light curve of classification error

将分类失败的数据进行整理和分析，结果表明分类错误主要来自以下两方面。（1）两个网站的光变曲线和分类信息不一致，如 AU Pup 和 AW Lac 两个目标星的原始数据与预处理后数据如图 6(a)，由图可知该光变曲线类型应为 EW 型，但 GCVS 给出的光变曲线类型均为 EB 型，可修改原始光变曲线类型数据消除这种错误。（2）光变曲线类型分类缺乏明确的区分标准，如图 6(b)，GCVS 给出 XZ Cmi 和 SW Lyn 分别为 EB 和 EA 型，但 CALEB 所给光变曲线数据非常接近，所以必需明确分类标准，并对原始数据逐条进行手工分类与核对，由于该工作量较大暂未进行。

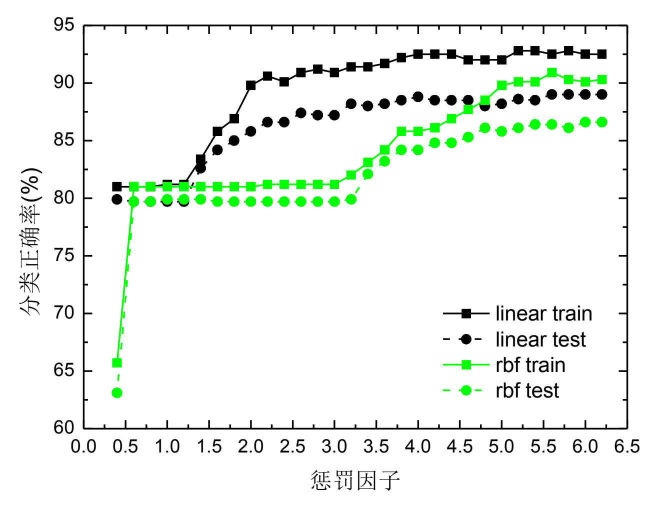


图 7 SVM 模型再次训练

修复原始数据分类错误 14 个目标，重新进行支持向量机模型训练和测试，结果如图 7。由于在上次的训练中已知由图可知 sigmoid 和 poly 效果较差，这次仅进行 linear 和 rbf 两种核函数的测试，由本次测试结果可知 linear 核函数结果较好，当惩罚因子设置为 5.8 时，能达到分类正确率为 92.8%（训练集）和 89.0%（测试集）。而如果使用 rbf 核函数，则惩罚因子设置为 5.6 时，能达到分类正确率为 90.9%（训练集）和 86.4%（测试集）。

随后准备 160 条光变曲线数据[13]，使用两种核函数训练好的模型进行测试，分类正确率均为 88.8%，检查了错误类型，主要是由于 EA 和 EB 两种光变曲线分类错误。

3. 总结和展望

本文提出一种基于机器学习的光变曲线自动分类算法，使用快速傅里叶变换提取目标数据的频率，选用偶次频率分量作为光变曲线特征值，使用所提取特征值训练支持向量机模型获得分类模型。随后采用 Python 编程实现上述算法并进行优化，实验数据使用 CALEB 的实测光变曲线数据和 GCVS 的分类数据，结果表明采用[f_0, f_2, f_4, f_6, f_8]作为特征值时，选用 linear 核函数，惩罚因子设置为 2.0 可获得最优分类结果，分类正确率为 89.8%（训练集）和 84.8%（测试集），能基本满足分类需求。

对分类错误数据进行分析，结果表明分类错误第 1 个原因来源于 CALEB 的光变曲线数据和 GCVS 分类信息不一致，该类错误可以通过修改分类信息消除。第 2 个原因来源于光变曲线类型分类缺乏明确区分标准，某些非常接近的光变曲线数据被分为不同类型，对最终测试结果造成干扰，需要制定明确的分类标准并对原始数据重新分类才可以避免该种错误。将第 1 种错误全部修正后，正确率提升到 92.8%（训练集）和 89.0%（测试集）。由于还未制定明确的分类标准，第 2 种错误来源还未修复。

在天文观测中自动化技术应用越来越广泛，而获取的数据量也越来越多，在常规的观测数据中往往包含着我们感兴趣的特殊数据，预示着特殊的天文现象如双星合并等，需要筛选特殊数据，然后对该目标进行重点观测，能获得更有用的数据结果。如何从大量数据中快速筛选出特殊数据是一个难点。在随后研究中将特殊光变曲线数据整理为样本数据，对支持向量机算法进行训练，使所得模型能够快速识别特殊光变曲线数据，从而能够快速响应。

参考文献

[1] 崔辰州, 于策, 肖健, 等. 大数据时代的天文学研究[J]. 科学通报, 2015,60(Z1):445-449.
Cui ChenZhou, Yu Ce, Xiao Jian, et al. Astronomy research in big-data era. Chin Sci Bull, 2015, 60(Z1):445-449

[2] 张海龙, 聂俊, 赵青, 等. 新疆天文台在线交叉证认服务[J]. 天文研究与技术, 2017,14(03):347-355.
Zhang Hailong, Nie Jun, Zhao Qing, et al. Xinjiang Astronomical Observatory Data Center Custom Uploading Crossmatcher [J]. Astronomical Research & Technology, 2017, 14(03):347-355.

[3] LEWIS H, RAFFI G. Advanced Software, Control, and Communication Systems for Astronomy[J]. Proc Spie, 2004,5496(1):65-68.

[4] 卫守林, 刘鹏翔, 王锋, 等. 基于Spark Streaming的明安图射电频谱日像仪实时数据处理[J]. 天文研究与技术, 2017,14(04):421-428.
Wei Shoulin, Liu Pengxiang, Wang Feng. Real-Time Data Processing in Mingantu Ultrawide Spectral Radio Heliograph Based on Spark Streaming [J]. Astronomical Research & Technology, 2017, 14(03):421-428.

[5] 陈淑鑫, 罗阿理, 孙伟民. R语言应用于LAMOST光谱分析初探[J]. 天文研究与技术, 2017,14(03):363-368.
Chen Shuxin, Luo Ali, Sun Weiming. Application of R language in LAMOST Spectral Analysis [J]. Astronomical Research & Technology, 2017, 14(03):363-368.

[6] ZHANG Y X, ZHOU X L, ZHAO Y H, et al. Statistical Study of 2XMMi-DR3/SDSS-DR8 Cross-correlation Sample[J]. Astronomical Journal, 2013,145(2):531-544.

[7] DEVINE T R, GOSEVAPOPSTOJANOVA K, MCLAUGHLIN M. Detection of dispersed radio pulses: a machine learning approach to candidate identification and classification[J]. Monthly Notices of the Royal Astronomical Society, 2016,459(2):w655.

[8] 杨远贵, 来春富. 密近双星的Roche势的计算[J]. 淮北师范大学学报(自然科学版), 2011,32(01):29-32.
Yang Yuangui, Lai Chunfu. Calcualting the Roche Potential of Close Binaries [J]. Journal of Huaibei Normal University (Natural Science). 2011,32(01):29-32.

[9] KIRK B, CONROY K, PRŠA A, et al. Kepler Eclipsing Binary Stars. VII. The Catalog of Eclipsing Binaries Found in the Entire Kepler Data-Set[J]. Astronomical Journal, 2016,151(3):68.

[10] POJMAŃSKI G. The All Sky Automated Survey[J]. Astronomische Nachrichten, 1997,325(6－8):467-481.

[11] AKERLOF C, AMROSE S, BALSANO R, et al. ROTSE All-Sky Surveys for Variable Stars. I. TestFields[J]. Astronomical Journal, 2000,119(4):1901.

[12] POJMANSKI G. The All Sky Automated Survey. Variable Stars in the 0h - 6h Quarter of the Southern Hemisphere[J]. Physics, 2002.

[13] https://www.researchgate.net/profile/Y-G_Yang.

An Automatic Classification Algorithm for Light Curve of Eclipsing Binary Stars based on SVM

Huiyu Yuan[1] Juan Zhao[2] Haifeng Dai[2] Yuangui Yang[2*]

1.Information College, HuaibeiNormal University, Huaibei 235000, China

2. HuaibeiNormal University Huaibei 235000, China

Abstract: This paper proposes an automatic classification algorithm for light curve of eclipsing binary stars based on machine learning. At first the algorithm normalizes the light curve and performs filtering/interpolatiing to reduce the noise effect in the preprocessing stage, then the Fourier coefficients, which are extracted by FFT from the light curve, are used as the feature vector to train SVM and a classification model is obtained. We implement this algorithm with Python and use the data captured from CALEB/GCVS to validate and discuss the effect of the feature vector, the SVM kernel function and the penalty coefficient to the classification accuracy. The correct rate of the classification model reaches 92.8% (training set) and 89% (test set). Finally, we use the third party data to verify the classification model and get a correct rate of 88.8%. The results prove the validity of the classification algorithm proposed in this paper.

Key words: automatic classification of optical curves; SVM; eclipsing binary stars